Cibersegurança e Educação: um estudo de caso com uso de IA para mitigação e detecção de ameaças contra crianças e adolescentes em ambientes digitais

Raquel Moreira Machado Fernandes [1, 2], Claudia Lage Rebello da Motta [2], Luiz Fernando Rust da Costa Carmo [2]

[1] Colégio Pedro II – Departamento de Informática Educativa – Campus São Cristóvão – Rio de Janeiro – Brasil

[2] Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal do Rio de Janeiro (UFRJ) – Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais Cidade Universitária – Ilha do Fundão – RJ – Brasil

raquel.fernandes.2@cp2.edu.br,{claudiam,rust}@nce.ufrj.br

Abstract. *This paper presents a thesis that investigates the exposure of children and adolescents to threats in digital environments, with the aim of mitigating impacts and supporting educational practices. The descriptive-interpretative methodology enabled the cataloging of 80 threats with the potential to affect the integrity of children and adolescents. A case study with students from Colégio Pedro II produced empirical evidence about their online experiences, contributing to the validation of the proposed taxonomy. As contributions, the proposal of the taxonomy, the construction of a Brazilian Portuguese dataset on cyberbullying, and the development of a prototype for automatic detection based on Artificial Intelligence are highlighted.*

Resumo. *Este artigo apresenta uma tese que investiga a exposição de crianças e adolescentes a ameaças em ambientes digitais, com o objetivo de mitigar impactos e subsidiar práticas educacionais. A metodologia, de abordagem descritivo-interpretativa, permitiu a catalogação de 80 ameaças com potencial de afetar a integridade de crianças e adolescentes. Um estudo de caso com estudantes do Colégio Pedro II produziu evidências empíricas sobre suas experiências online, contribuindo para a validação da taxonomia proposta. Como contribuições, destacam-se a proposição da taxonomia, a construção de um dataset em português brasileiro sobre cyberbullying e o desenvolvimento de um protótipo de detecção automática baseado em Inteligência Artificial.*

1. Introdução

Em 2020, a pandemia de *Sars-CoV-2* provocou mudanças abruptas na relação da sociedade com a tecnologia, evidenciando desigualdades de acesso e novas formas de vulnerabilidade. Nesse período, observou-se um aumento expressivo de ocorrências de segurança cibernética (Check Point Research, 2021; Kaspersky, 2020; CISO Advisor, 2022), as quais também alcançaram crianças e adolescentes (Instituto Alana, 2021; SaferNet, 2023), trazendo à tona ameaças como cyberbullying, *deep nude*, exploração sexual e exposição indevida de dados pessoais, entre outros.

Nesse contexto, identifica-se como problema a ausência de um mapeamento sistemático de ameaças em ambientes digitais direcionadas a crianças e adolescentes, capaz de orientar o desenvolvimento de estratégias e soluções de prevenção, detecção e mitigação. Tal mapeamento também se mostra necessário para apoiar a compreensão do cenário por pais, responsáveis e profissionais da educação, como professores e

pedagogos, em geral sem formação específica em Computação, favorecendo sua atuação em ações pedagógicas, preventivas e de denúncia.

O objetivo desta pesquisa, portanto, foi investigar a exposição de crianças e adolescentes a ameaças em ambientes digitais, a partir das seguintes hipóteses: 1) a de que é possível elaborar uma taxonomia capaz de organizar e classificar essas ameaças, de forma estruturada, com base em um vocabulário controlado; e 2) a de que o uso de Inteligência Artificial (IA) pode auxiliar na detecção automática dessas ameaças, por meio de técnicas de aprendizado de máquina e Processamento de Linguagem Natural (PLN), contribuindo para o desenvolvimento de soluções computacionais de apoio a estratégias educacionais, institucionais e tecnológicas de mitigação dessas ameaças.

2. Metodologia

Esta pesquisa é de natureza básica, com abordagem qualitativa de caráter descritivo-interpretativo. Inicialmente, por meio de uma pesquisa exploratória, elaborou-se uma proposta preliminar de taxonomia com 45 ameaças (Fernandes et al., 2022). Posteriormente, a taxonomia foi revista, traduzida e ampliada utilizando-se como procedimento técnico a pesquisa bibliográfica na perspectiva de Prodanov e Freitas (2013) com objetivo ampliar a taxonomia, incluindo: 1) ameaças não identificadas a partir das palavras-chave selecionadas na fase anterior; 2) ameaças não identificadas a partir das bases de dados selecionadas na fase anterior; e 3) novas ameaças que ainda não existiam, não tinham sido relatadas ou não eram populares no período anterior. Ao término desta etapa, foram catalogadas ao todo 80 ameaças.

A pesquisa também apoia-se em um estudo de caso conduzido no Colégio Pedro II que permitiu ouvir diretamente 133 estudantes do 4º ano do ensino fundamental sobre suas experiências em ambientes digitais. Aprovado pelo Comitê de Ética, o estudo contemplou evidências qualitativas e quantitativas na análise das ameaças a partir da experiência dos estudantes, onde o *cyberbullying* se mostrou uma ameaça significativa juntamente com *Phishing*, sobretudo associado ao Roblox, bem como conteúdos prejudiciais, publicidade infantil e *malwares*.

Nesta etapa da pesquisa, foram investigadas 45 ameaças com base em Fernandes et al. (2022) e verificou-se que nenhuma delas apresentou ausência de ocorrências, isto é, todas foram mencionadas por pelo menos 3% da amostra. Embora esses dados não permitam confirmar a ocorrência objetiva dos eventos, indicam um nível mínimo de exposição em todas as ameaças analisadas, dado que merece atenção da sociedade.

Após a delimitação do escopo, foi conduzida a etapa aplicada da pesquisa, que envolveu o desenvolvimento e a validação de um protótipo funcional de detecção automática do *cyberbullying*. Para tanto, contemplou-se as plataformas mais utilizadas pela amostra (Youtube, TikTok e Instagram) e coletou-se comentários de vídeos selecionados a partir dos interesses informados pela amostra para criação de um *dataset* sobre *cyberbullying*. Tal coleta foi realizada em conformidade com as diretrizes das plataformas utilizadas e com boas práticas reconhecidas em pesquisas de Ciência de Dados e Análise de Redes Sociais (ARS), incluindo a extração apenas de comentários públicos, a anonimização de nomes de usuários e identificadores, a preservação da privacidade dos indivíduos envolvidos, entre outras. Após limpeza e preparação, o

processo de anotação foi conduzido manualmente com base em diretrizes de anotação propostas nesta pesquisa especificamente para *cyberbullying* em língua portuguesa no âmbito de crianças e adolescentes no contexto brasileiro.

O dataset elaborado foi utilizado para treinamento de modelos de *machine learning*, uma subárea da IA. Foram aplicados dois algoritmos amplamente utilizados em problemas de detecção textual: *Support Vector Machine* (SVM) e *Naive Bayes* (NB). Os dados foram divididos em conjuntos de treino (80%) e teste (20%), com base em Eberhart, Ignaczak e Martins (2021). Para o balanceamento dos dados selecionou-se a técnica de *Data Augmentation* (Sun et al., 2020; Guo et al., 2020), que foi implementada em Python através da biblioteca Pandas.

3. Resultados

Esta pesquisa, de relevância social e educacional, com contribuições teóricas e práticas cumpriu seus objetivos ao mapear 80 ameaças digitais enfrentadas por crianças e adolescentes na atualidade, além de propor uma solução de suporte à detecção de *cyberbullying* com uso de IA, validando as hipóteses consideradas. O *dataset* contendo 52,092 comentários, sendo 26,046 anotados como *cyberbullying* e 26,046 *n_cyberbullying,* tem potencial educacional e pode servir em pesquisas futuras na área de Computação. Para fins de reprodutibilidade técnica, foi disponibilizado em: https://www.kaggle.com/datasets/raquelfernandes2024/cyberbullying-br.

Em relação à detecção, cujo protótipo está disponível em https://cyberbullying-br.streamlit.app/ e o código-fonte em: https://github.com/raquelmachado4993/cyberbullying-br, o modelo treinado com SVM apresentou acurácia de 96,8%, precisão de 98,2%, recall de 95,5% e F1-Score de 96,8% na tarefa de detecção de *cyberbullying*, enquanto o modelo treinado com NB obteve acurácia de 96,5%, precisão de 98,3%, recall de 94,6% e F1-Score de 96,4%. Ambos os modelos apresentaram desempenho satisfatório e resultados próximos. Foram utilizadas sentenças contrafactuais (Gardner et al., 2020; Qin et al., 2019) como estratégia para minimizar o *overfitting*. Ainda assim, os elevados índices de desempenho sugerem prudência ao se inferir a total superação deste problema. Portanto, recomenda-se, em trabalhos futuros, a aplicação de técnicas de validação mais robustas como *K-fold Cross Validation*. Além disso, podem ser exploradas estratégias complementares de classificação e detecção, incluindo o uso de ontologias, modelagem semântica e análise de sentimentos, entre outras. Sugere-se também aprimoramento e ampliação do *dataset.*

Outras limitações da pesquisa incluem o recorte metodológico adotado na construção da taxonomia e, no estudo de caso, a utilização de uma única amostra, além do uso de instrumentos de autorrelato, entre outras.

Ainda que o uso de soluções computacionais, sobretudo com IA, seja promissor, o desenvolvimento do pensamento crítico é essencial para formar crianças e adolescentes para o uso ético, seguro e responsável das tecnologias. Esta pesquisa fornece um referencial claro para reconhecer e compreender as ameaças, contribuindo para a Educação em Computação e para o empoderamento infantojuvenil frente aos desafios dos ambientes digitais em consonância com a Base Nacional Comum Curricular (BNCC).

Referências

Brasil. Ministério da Educação (2022) “Parecer CNE/CEB nº 2, de 17 de fevereiro de 2022. Normas sobre Computação na Educação Básica – Complemento à Base Nacional Comum Curricular (BNCC)”.

Check Point Research (2021) “Cyber Security Report 2021”, https://www.checkpoint.com/pt/pages/cyber-security-report-2021/

CISO Advisor (2022) “Ciberataques ao setor de educação no Brasil crescem 122%”, https://www.cisoadvisor.com.br/ciberataques-ao-setor-de-educacao-no-brasil-crescem-122/.

Eberhart, C; Ignaczak, L; Martins, M. (2021) “Text Mining for Cyberbullying Detection: a Brazilian Portuguese Evaluation”, In: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 13., 2021. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 92–100.

Fernandes, R., Carmo, L. and Motta, C. (2022). A taxonomy proposal of cyber threats involving children and adolescents. XVII Latin American Conference on Learning Objects and Technology (LACLO), https://ieeexplore.ieee.org/document/10013466.

Fernandes, R. (2025) “Cibersegurança e Educação: Um estudo de caso com uso de inteligência artificial para mitigação e detecção de ameaças contra crianças e adolescentes em ambientes digitais”, Rio de Janeiro, 246 f.

Gardner, M.; Artzi, Y.; Basmov, V.; Berant, J.; Bogin, B.; Chen, S.; Dasigi, P.; Dua, D.; Elazar, Y.; Gottumukkala, A.; Gupta, N.; Hajishirzi, H.; Ilharco, G.; Khashabi, D.; Lin, K.; Liu, J.; Liu, N. F.; Mulcaire, P.; Ning, Q.; Singh, S.; Smith, N. A.; Subramanian, S.; Tsarfaty, R.; Wallace, E.; Zhang, A.; Zhou, B. (2020) “Evaluating models’ local decision boundaries via contrast sets”, In: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

Guo, D; Kim, Y.; Rush, A. (2020) “Sequence-Level Mixed Sample Data Augmentation”, In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 5547–5552.

Instituto Alana; InternetLab (2021) “O direito das crianças à privacidade: obstáculos e agendas de proteção à privacidade e ao desenvolvimento da autodeterminação informacional das crianças no Brasil. Contribuição conjunta para o relator especial sobre o direito à privacidade da ONU”, https://internetlab.org.br/wp-content/uploads/2021/03/ilab-alana_criancas-privacidade_PT_20210214-4.pdf.

Kaspersky (2020) “Brasileiros estão entre os mais atacados por golpes durante a pandemia”, https://www.kaspersky.com.br/blog/brasil-phishing-covid-golpe/15902/

Prodanov, C.; Freitas, E. (2013) “Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico”, Novo Hamburgo, RS: Feevale, 2013.

SaferNet Brasil (2023) “Denúncias de imagens de abuso e exploração sexual infantil online compartilhadas pela SaferNet com as autoridades têm aumento de 70% em 2023”,

https://new.safernet.org.br/content/denuncias-de-imagens-de-abuso-e-exploracao-sexual-infantil-online-compartilhadas-pela

Qin, L.; Bosselut, A.; Holtzman, A.; Bhagavatula, C.; Clark, E.; Choi, Y. (2019) "Counterfactual story reasoning and generation", In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, Hong Kong, China. Association for Computational Linguistics, p. 5043–5053.

Sun, L.; Xia, C; Yin, W.; Liang, T.; Yu, S. Philip; He, L. (2020) "MixupTransformer: Dynamic data augmentation for NLP tasks", In: Proceedings of the 28th International Conference on Computational Linguistics, 2020. p. 3436–3440.